Cauchyn jakauma ja sen mahdollisuus lukio-opetuksessa

Jasmiina Rintala, Helsingin yliopisto

7.5.2020

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

MATEMAATTIS-LUONNONTIETEELLINEN TIEDEKUNTA
MATEMATISK-NATURVETENSKAPLIGA FAKULTETEN
FACULTY OF SCIENCE

Tiedekunta – Fakultet – Faculty	Koulutusohjelma – Utbildningsprogram – Degree programme
Matemaattis-luonnontieteellinen	Matematiikan opettajan maisteriohjelma

Tekijä – Författare – Author

Jasmiina Rintala

Työn nimi – Arbetets titel – Title

Cauchyn jakauma ja sen mahdollisuus lukio-opetuksessa

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Pro gradu -tutkielma	Toukokuu 2020	39

Tiivistelmä – Referat – Abstract

Tämä tutkielma käsittelee Cauchyn jakaumaa ja siitä muunneltua log-Cauchyn jakaumaa. Cauchyn jakauma on jatkuva ja todella paksuhäntäinen ja sallii siksi poikkeamia aineistossa, jonka takia se on potentiaalinen vaihtoehto erilaisten luonnonilmiöiden mallintamisessa.

Käyn ensimmäisessä luvussa läpi, mikä standardi Cauhcyn jakauma on: mitä matemaattisia määritelmiä sen johtamiseen tarvitaan ja kuinka se johdetaan. Tutkielmassa todistetaan, että tällä jakaumalla ei ole olemassa odotusarvoa eikä varianssia. Puolestaan moodi ja mediaani voidaan laskea ja huomataankin, että ne ovat Cauchyn jakaumalla samat.

Käsittelen lyhyesti logaritmisen Cauchyn jakauman ja johdan sen tiheys- ja kertymäfunktiot. Tämän jälkeen perehdyn sekä Cauchyn että log-Cauchyn jakaumien erilaisiin sovelluksiin. Jotta lukija saa käsityksen jakaumien käyttötarkoituksista, käy läpi useita tutkimuksia kevyesti. Muutamassa tutkimuksessa huomataan, että Cauchyn ja log-Cauchyn jakauma sopivat kyseisiin mallinnuksiin hyvin.

Viimeisessä osiossa pohdin Cauchyn jakauman mahdollisuuksia lukio-opetuksessa uusimman lukion opetussuunnitelman (2019) pohjalta. Esitän lopuksi oman ehdotukseni projektityöstä pitkän matematiikan kurssille MAA12 ja perustelen sen sopivuutta kyseiselle valinnaiselle kurssille. Tämä projektityö kehittää oppilaan laaja-alaista osaamista ja luo hyvän kokonaisuuden oppiainerajat ylittävään opetukseen.

Avainsanat – Nyckelord – Keywords

Cauchyn jakauma, log-Cauchyn jakauma

Säilytyspaikka – Förvaringställe – Where deposited

Kumpulan kampuskirjasto

Muita tietoja – Övriga uppgifter – Additional information

Sisältö

Termistöä

Arvojoukko: perusjoukko, jossa satunnaisilmiön sijasta käytetään satunnaismuuttujan arvoja.

Diskreetti satunnaismuuttuja: muuttujan arvoalue muodostuu erillisistä reaaliakselin pisteistä. Arvoalue on siis aina joko äärellinen tai korkeintaan numeroituvasti ääretön.

Jatkuva satunnaismuuttuja: muuttujan arvoalue muodostuu jostain reaaliakselin osavälistä ja on täten ylinumeroituva.

Kvantiili: osuuspisteitä eli *faktiileja*, jotka jakavat suuruusjärjestykseen asetetun muuttujan jakauman tietyn suuruisiin osiin. Yleisimpiä kvantiileja ovat mediaani, kvartiilit, kvintiilit, desiilit ja persentiilit.

Momentti: Satunnaismuuttujan jakaumasta määritelty *tunnusluku*, jonka avulla jakaumaa voidaan luonnehtia. Momentit määritellään odotusarvon avulla.

Perusjoukko eli otosavaruus: satunnaisilmiön kaikkien erilaisten alkeistapausten joukko.

Satunnaismuuttuja: sellainen reaaliarvoinen funktio, joka liittää tietyn tapahtuman alkeistapaukseen numeerisen arvon, eli reaaliluvun.

Todennäköisyysjakauma: muodostuu satunnaismuuttujan arvoista ja niihin liitetyistä todennäköisyyksistä. Todennäköisyysjakauma kuvaa, kuinka yleisiä satunnaismuuttujan eri arvot ovat.

Typistetty jakauma: sellainen jakauma, joka on skaalattu siten, että sitä voi tarkastella vain tietyllä välillä, esimerkiksi $[0, \infty]$.

1 Johdanto

Cauchyn jakauma on saanut nimensä lahjakkaan ranskalaisen matemaatikon *Augustin Louis Cauchyn* mukaan. Hän oli yksi analyysin pioneereista ja tutki mm. permutaatioryhmiä. Cauchyn jakaumaa nimitetään usein myös *Lorentzin jakaumaksi* (Hendrik Lorentzin mukaan), *Cauchy-Lorentzin jakaumaksi* tai *Breitin-Wignerin jakaumaksi*.

Cauchyn jakauma on jatkuva todennäköisyysjakauma, joka muistuttaa paljon normaalijakaumaa. Se on kuitenkin ns. *paksuhäntäinen jakauma*, jossa kuvaajan häntien alla on normaalijakaumaa enemmän todennäköisyysmassaa (kts. Kuva 1). Tällä jakaumalla ei ole varianssia eikä odotusarvoa, mitkä ovat tyypillisiä tunnuslukuja todennäköisyysjakaumille. Jakaumalla ei myöskään ole muita äärellisiä *momentteja*, mutta jotkin korkeammista momenteistaista voidaan määrittää. Näiden sijaan Cauchyn jakaumalla on hyvin määritelty moodi ja mediaani, sekä mediaanin suhteen symmetrinen kuvaaja. Sitä voidaan käyttää joustavissa menetelmissä, sillä se sallii satunnaisotoksissa isotkin poikkeamat.

Toisessa luvussa johdan erikseen Cauchyn jakauman tiheys- ja kertymäfunktiot. Geogebraa apuna käyttäen olen piirtänyt molemmista funkioista kuvan havainnollistamaan niiden kulkua xy-akselilla. Tämän jälkeen käsittelen jakauman odotusarvoa ja varianssia, joita tällä ei tunnestusti ole olemassa. Lisäksi määritän jakauman moodin ja mediaanin.

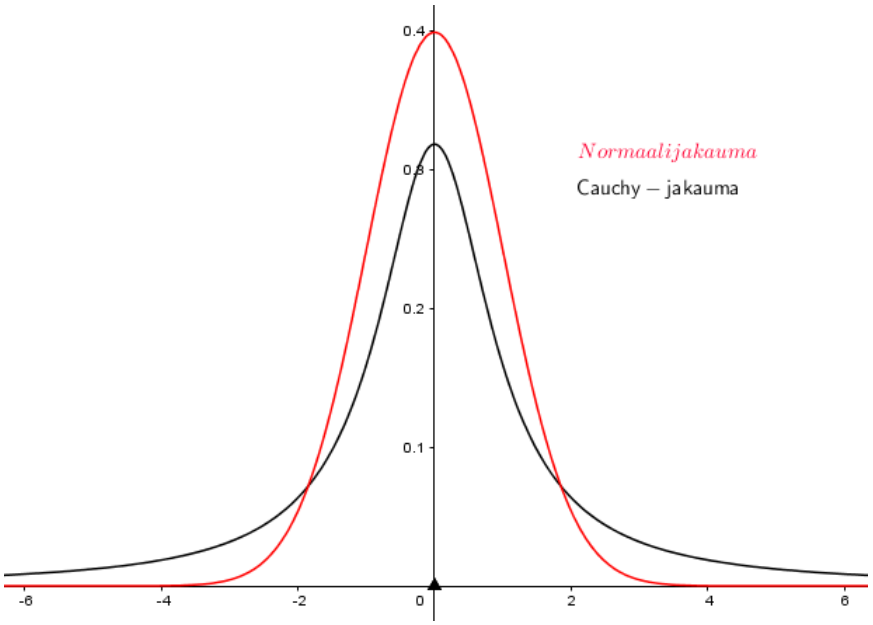
Kolmannessa luvussa käyn läpi, miltä yleinen Cauchyn jakauman tiheys- ja kertymäfunktiot näyttävät eri parametrien arvoilla. Näitä olen havainnollistanut parilla kuvalla, joissa annan parametreille vain hieman eri arvoja. Näin nähdään helposti, miten ne vaikuttavat funktioiden kulkuun. Neljännessä luvussa annan muutamia esimerkkejä jakauman yhteyksistä muihin jakaumiin.

Seuraavana käsittelen kevyesti logaritmisen Cauchyn jakauman. Johdan sen tiheys- ja kertymäfunktiot, mutta jakauman momenttien käsittelyn jätän pois log-Cauchyn osalta. Pääpaino loppuosassa on jakaumien konkreettisissa käyttöesimerkeissä. Käsittelen luvussa 6 Cauchyn jakauman käyttöä sademäärien mallinnuksissa. Käyn läpi kahden eri tutkimuksen idean ja tulokset lyhyesti läpi. Ensimmäisessä tutkimuksissa, joita artikkelini käsittelevät, tullaan huomaamaan, että Cauchyn jakauma on sopiva malli Sri-Lankan sademäärän enimmäismäärien kuvaamiseen. Toisessa tutkimuksessa jakauma on yksi vaihtoehdoista, mutta ei sovi parhaiten mallintamaan tämän tutkimuksen aineistoja.

Luvussa 7 käyn läpi log-Cauchyn jakauman sovelluksia. Ensimmäisessä osiossa käsittelen artikkelia, joka liittyy HI-viruksen itämis- ja tartunta-aikaan. Jakauma on mielenkiintoinen kandidaatti paksuhäitäisyytensä takia, mutta sitä ei kuitenkaan todeta sopivaksi malliksi. Toisessa osiossa käyn läpi ekologiaan liittyviä artikkeleita, joissa tarkoituksena on löytää sopiva jakauma kuvaamaan lajirunsauksia eri ekologisilla alueilla. Log-Cauchyn jakauma on edolla jokaisessa ja todetaan myös

sopivaksi malliksi kaikissa kolmessa tutkimuksessa, yhdessä jopa parhaimmaksi.

Viimeinen luku käsittelee Cauchyn jakauman mahdollisuuksia lukio-opetuksessa. Kerron oman ehdotukseni projektityöstä, jonka voisi toteuttaa pitkän matematiikan kurssilla **MAA12 Analyysi ja jatkuva jakauma (2 op)**. Pohdin lukion opetussuunnitelman 2019 pohjalta, kuinka kyseinen projekti sopisi kurssille.



Kuva 1: Normaali- ja Cauchy-jakauma

2 Cauchyn jakauma

2.1 Tiheysfunktio

Tiheysfunktio (engl. probability density function) kertoo jatkuvan satunnaismuuttujan todennäköisyysjakauman muodon. Koska kyseessä on jatkuva satunnaismuuttuja, sen arvojoukko on ylinumeroituva eikä todennäköisyyksiä tiheysfunktiosta voida määritellä pisteittäin. Tiheysfunktiolla f on seuraavat ominaisuudet

1. $f(x) \geq 0$, kaikilla $x \in \mathbb{R}$,
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$,
3. $P(a < X < b) = \int_a^b f(x)\,dx$,
4. $P(X = x) = 0$, kaikilla $x \in \mathbb{R}$.

Kohdassa neljä viitataan siihen, että minkä tahansa jatkuvan jakauman *yksittäisen* tapahtuman todennäköisyys on nolla. Diskreetin jakauman tapauksessa tilanne on toinen, kun satunnaismuuttujan todennäköisyydet voidaan määritellä pisteittäin. Jotta funktio on todella tiheysfunktio, sen tulee toteuttaa kaksi ensimmäistä ominaisuutta.

Käsitellään seuraavaksi Cauchyn jakaumaa ja määritellään sen tiheysfunktio.

Lause 2.1. Funktio

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, \quad x \in \mathbb{R} \tag{1}$$

määrittelee jatkuvan jakauman tiheysfunktion. Tätä jakaumaa kutsutaan standardiksi Cauchyn jakaumaksi.

Todistus.

$$\text{Selvästi } f(x) > 0 \qquad \text{kaikilla } x \in \mathbb{R}.$$

Jotta voidaan tutkia funktion f integraalia, täytyy tietää kuinka arkusfunktiot käyttäytyvät.
Olkoon $y = \arctan(x) \Leftrightarrow x = \tan(y)$. Nyt y on määritelty kaikilla $x \in \mathbb{R}$. Tiedetään, että

$$\tan(y) \in \mathbb{R} \Leftrightarrow -\frac{\pi}{2} < y < \frac{\pi}{2}, \quad \text{joten}$$

$$\lim_{x\to\infty} \arctan(x) = \frac{\pi}{2} \quad \text{ja} \quad \lim_{x\to-\infty} \arctan(x) = -\frac{\pi}{2}.$$

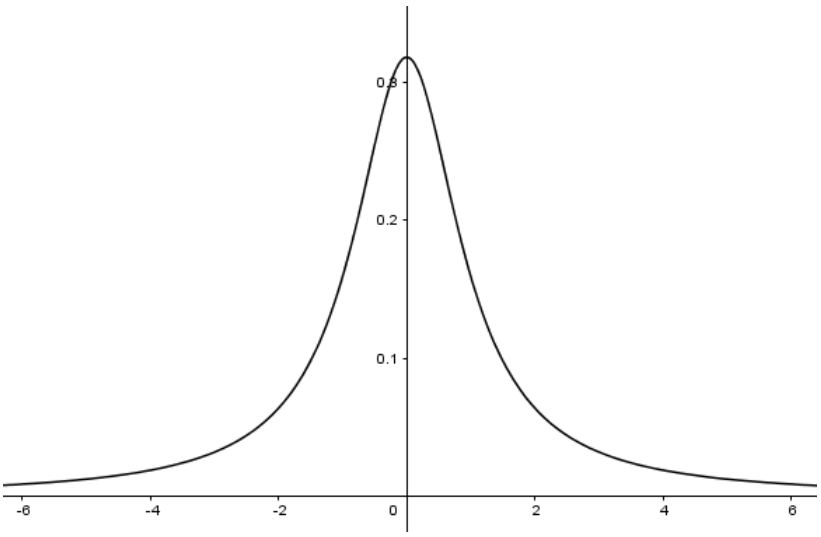
Hyödynnetään lisäksi käänteisfunktion deirvoimissääntöä: kun $y = \tan(x)$, niin

$$\begin{aligned} D\tan(y) &= \frac{1}{\cos^2(y)} = 1 + \tan^2(y), \text{ joten} \\ D\arctan(x) &= \frac{1}{D\tan(y)} = \frac{1}{1+\tan^2(y)} = \frac{1}{1+x^2}. \end{aligned}$$

Näiden tulosten perusteella saadaan, että

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,dx &= \frac{1}{\pi} \lim_{a\to-\infty} \lim_{b\to\infty} \int_a^b \frac{1}{1+x^2}\,dx \\ &= \frac{1}{\pi} \lim_{a\to-\infty} \lim_{b\to\infty} [\arctan(x)]_a^b \\ &= 1. \end{aligned}$$

Koska tiheysfunktio integroituu ykköseksi ja on aidosti positiivista kaikilla $x \in \mathbb{R}$, niin se todellakin on tiheysfunktio. □



Kuva 2: Cauchy-jakauman tiheysfunktio

2.2 Kertymäfunktio

Kertymäfunktio (engl. cumulative distribution function) $F(x)$ kuvaa satunnaismuuttujan X todennäköisyyden jakautumista. Kertymäfunktion saama arvo x

on todennäköisyys sille, että satunnaismuuttuja X saa korkeintaan arvon x, eli $F(x) = P(X \leq x)$.

Jatkuvan jakauman tapauksessa kertymäfunktio saadaan tiheysfunktiota integroimalla, ts. jos $f(x)$ on satunnaismuuttujan X tiheysfunktio, niin $F : \mathbb{R} \to [0, 1]$ on

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt\,. \tag{2}$$

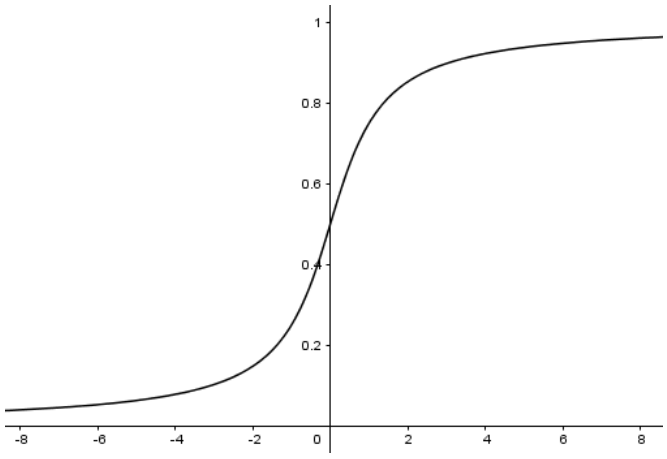
Lause 2.2. Cauchyn jakauman kertymäfunktio on

$$F(x) = \frac{\arctan(x)}{\pi} + \frac{1}{2}\,. \tag{3}$$

Todistus. Olkoon $f(x)$ Cauchyn jakauman tiheysfunktio. Tällöin

$$\begin{aligned} F(x) = \int_{-\infty}^{x} f(x)\, dx &= \frac{1}{\pi} \int_{-\infty}^{x} \frac{1}{1+x^2}\, dx \\ &= \frac{1}{\pi} \arctan(x)\Big|_{-\infty}^{x} \\ &= \frac{1}{\pi} \left(\arctan(x) - \arctan(-\infty)\right) \\ &= \frac{1}{\pi} \arctan(x) - \frac{1}{\pi}(-\frac{1}{2}\,\pi) \\ &= \frac{1}{\pi} \arctan(x) + \frac{1}{2} \\ &= \frac{\arctan(x)}{\pi} + \frac{1}{2}\,. \end{aligned}$$

□



Kuva 3: Cauchy-jakauman kertymäfunktio

2.3 Odotusarvo ja varianssi

Odotusarvo (engl. expected value) on todennäköisyyslaskennassa satunnaismuuttujan odotettavissa oleva arvo ja todennäköisyysjakauman ensimmäinen tunnusluku. *Varianssi* (engl. variance) puolestaan kuvaa satunnaismuuttujan hajonnan suuruutta, eli kuinka paljon satunnaismuuttujan arvot vaihtelevat odotusarvosta. Varianssin σ^2 määrittämiseen tarvitaan odotusarvo.

Jatkuvalle jakaumalle odotusarvo määritetään hieman eri tavalla kuin diskreetille jakaumalle. Seuraavat määritelmät tarvitaan Cauchyn jakauman odotusarvon ja varianssin todistamiseen.

Määritelmä 2.1. Olkoon X jatkuva satunnaismuuttuja, jonka tiheysfunktio on f. Odotusarvo satunnaismuuttujalle X on luku

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx \;<\; \infty\,, \tag{4}$$

edellyttäen, että integraali suppenee itseisesti, eli $E|X| = \int_{-\infty}^{\infty} |x|f(x)\,dx \;<\; \infty$. Jos integraali ei ole itseisesti suppeneva, niin sanotaan, että satunnaismuuttujalla X *ei* ole odotusarvoa.

Määritelmä 2.2. Olkoon X satunnaismuuttuja, jolla on odotusarvo $\mu = E(X)$. Tällöin X:n varianssi on

$$\sigma^2(X) = E((X-\mu)^2)\,, \tag{5}$$

edellyttäen, että kyseinen odotusarvo on olemassa.

Lause 2.3. Cauchyn jaukaumalla ei ole odotusarvoa, eikä täten myöskään varianssia.

Todistus. Olkoon $X \sim \text{Cauchy}(0,1)$ ja $f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$ sen tiheysfunktio. Tällöin

$$\begin{aligned} E|X| &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{|x|}{1+x^2}\,dx \\ &= \frac{2}{\pi}\int_{0}^{\infty} \frac{x}{1+x^2}\,dx \\ &= \frac{1}{\pi}\left[\ln(1+x^2)\right]_0^{\infty} \\ &= \infty. \end{aligned}$$

Koska $E|X| = \infty$, niin sanotaan, että odotusarvoa ei ole olemassa.

Nyt

$$\sigma^2 = E((X-\mu))^2),$$

missä $\mu = E(X)$ on satunnaismuuttujan X odotusarvo. Koska odotusarvoa ei ole olemassa, varianssia ei voida laskea. Täten Cauchyn jakaumalla ei ole olemassa odotusarvoa, eikä varianssia. □

Myös muiden korkeampien momenttien laskemisessa hyödynnetään jakauman odotusarvoa. Koska sitä ei ole, voidaan päätellä ettei Cauchyn jakaumalla ole korkeampia momentteja.

2.4 Moodi ja mediaani

2.4.1 Moodi

Moodi (engl. mode) eli tyyppiarvo on se satunnaismuuttujan arvo, jonka frekvenssi on suurin, ts. se arvo, joka esiintyy aineistossa useimmin.

Määritelmä 2.3. Olkoon X jatkuva satunnaismuuttuja, jonka tiheysfunktio on f. Tällöin jakauman *moodi* on tiheysfunktion maksimikohta, eli kohta, missä se saa maksimiarvonsa. Ts. jos x_0, $x \in \mathbb{R}$ siten, että

$$f(x_0) = \max f(x),$$

niin x_0 on satunnaismuuttujan X moodi.

Jotta voimme tarkastella funktion maksimikohtaa, tarvitaan muutama aputulos funktion monotonisuudesta ja derivaatasta.

Määritelmä 2.4. Olkoon $A \subset \mathbb{R}$. Funktio $f : A \to \mathbb{R}$ on joukossa A

- aidosti kasvava, jos $f(a) > f(b)$, kun $a > b$,
- aidosti vähenevä, jos $f(a) < f(b)$, kun $a > b$ kaikilla $a, b \in A$.

Lemma 2.4.1. Olkoon $f : (a, b) \to \mathbb{R}$ jatkuva ja derivoituva funktio ja f' sen derivaattafunktio. Tällöin, jos

$$\begin{aligned} f'(x) > 0 \quad &\text{kaikilla } x \in (a,b) \quad \Rightarrow \quad f \text{ on aidosti kasvava välillä } (a,b) \\ f'(x) < 0 \quad &\text{kaikilla } x \in (a,b) \quad \Rightarrow \quad f \text{ on aidosti vähenevä välillä } (a,b). \end{aligned}$$

Määritetään seuraavaksi Cauchyn jakauman moodi, eli tässä tapauksessa sen tiheysfunktion maksimikohta.

Lause 2.4. Cauchyn jakauman moodi on pisteessä $x = 0$.

Todistus. Olkoon $X \sim \text{Cauchy}(0, 1)$ ja $f(x) = \frac{1}{\pi(1+x^2)}$ jakauman tiheysfunktio. Tällöin

$$\begin{aligned} f'(x) &= f'(\pi(1+x^2)^{-1}) \\ &= -1(\pi(1+x^2))^{-2}\, 2\pi x \\ &= -\frac{2\pi x}{(\pi(1+x^2))^2} \\ &= -\frac{2x}{\pi(1+x^2)^2}\,. \end{aligned}$$

Lasketaan seuraavaksi derivaatan nollakohdat. $f'(x) = 0$, kun

$$\begin{aligned} -\frac{2x}{\pi(1+x^2)^2} &= 0 \\ \Leftrightarrow\ 2x &= 0 \\ \Leftrightarrow\ x &= 0\,. \end{aligned}$$

Koska tiheysfunktion derivaattafunktiolla on täsmälleen yksi nollakohta, tiedetään, että tiheysfunktiolla on täsmälleen yksi maksimiarvo. Tutkitaan vielä tiheysfunktion monotonisuutta derivaatan avulla.
Oletetaan, että $x < 0$. Tällöin

$$\frac{-2x}{\pi(1+x^2)^2} > 0, \qquad \text{sillä } -2x > 0 \quad \text{ja} \quad \pi(1+x^2)^2 > 0.$$

Oletetaan, että $x > 0$. Tällöin

$$\frac{-2x}{\pi(1+x^2)^2} < 0, \qquad \text{sillä } -2x < 0 \quad \text{ja} \quad \pi(1+x^2)^2 > 0.$$

Siis f on aidosti kasvava, kun $x \in (-\infty, 0)$ ja aidosti vähenevä, kun $x \in (0, \infty)$. Lisäksi $f'(0) = 0$, joten standardin Cauchyn jakauman maksimikohta eli moodi on kohdassa $x = 0$. □

2.4.2 Mediaani

Kun aineiston havaintoarvot laitetaan suuruusjärjestykseen, niiden keskimmäistä arvoa sanotaan *mediaaniksi* (engl. median). Jos havaintoarvoja on parillinen määrä, mediaani määritetään kahden keskimmäisen arvon keskiarvosta. Tämä tilanne on kuitenkin mahdollinen vain diskreettien satunnaismuuttujien yhteydessä.

Seuraava määritelmä on olennainen jatkuvan Cauchyn jakauman mediaania määritettäessä.

Määritelmä 2.5. Olkoon X jakuva satunnaismuuttuja ja $m \in \mathbb{R}$. Tällöin jakauman X *mediaani* on luku m, jolle pätee

$$P(X \leqslant m) \geqslant \frac{1}{2} \qquad \text{ja} \qquad P(X \geqslant m) \geqslant \frac{1}{2}\,. \tag{6}$$

Lause 2.5. Standardin Cauchyn jakauman mediaani on kohdassa $x_0 = 0$.

Todistus. Olkoon $X \sim \text{Cauchy}(0,1)$ ja $F(x) = \frac{\arctan(x)}{\pi} + \frac{1}{2}$ jakauman kertymäfunktio. Tutkitaan ensin tilannetta $P(X \leqslant x_0)$, jossa $x_0 = \frac{1}{2}$. Nyt

$$\lim_{x_0 \to \frac{1}{2}^+} \frac{\arctan(x_0)}{\pi} + \frac{1}{2} = \frac{\arctan(\frac{1}{2})}{\pi} + \frac{1}{2} = 0 + \frac{1}{2} = \frac{1}{2}.$$

Seuraavaksi tilanne $P(X \geqslant x_0)$. Tällöin

$$\lim_{x_0 \to \frac{1}{2}^-} \frac{\arctan(x_0)}{\pi} + \frac{1}{2} = \frac{\arctan(\frac{1}{2})}{\pi} + \frac{1}{2} = 0 + \frac{1}{2} = \frac{1}{2}.$$

Siis $P(X \leqslant \frac{1}{2}) = \frac{1}{2}$ ja $P(X \geqslant \frac{1}{2}) = \frac{1}{2}$. Täten kohta $x_0 = 0$ on standardin Cauchyn jakauman mediaani. □

Tulosten perusteella huomataan, että standardilla Cauchyn jakaumalla moodi ja mediaani ovat samassa pisteessä $x = 0$.

3 Yleinen Cauchyn jakauma

3.1 Tiheysfunktio

Yleisen Cauchyn jakauman tiheysfunktio on

$$f(x;\theta,\sigma) = \frac{1}{\pi\sigma\left[1+\left(\frac{x-\theta}{\sigma}\right)^2\right]} = \frac{1}{\pi\sigma}\left[\frac{\sigma^2}{(x-\theta)^2+\sigma^2}\right], \tag{7}$$

jossa $\theta \in \mathbb{R}$ on funktion *sijaintiparametri* ja $\sigma > 0$ sen *skaalaparametri*. Sijaintiparametri kertoo, missä kohtaa tiheysfunktion huippu on. Skaalaparametri puolestaan kertoo tiheysfunktion huipusta: mitä suurempi σ on, sitä leveämpi huippu on.

Aiemmin huomattiin, että tiheysfunktion huippu on kohdassa $x = \theta$, jolloin tiheysfunktion maksimiarvo on

$$\max f(x) = f(\theta) = \frac{1}{\pi\sigma}.$$

Cauchyn jakauman tiheysfunktion ominaisuuksia

1. Cauchyn jakauma on tiheysfunktio $f(x)$ on positiivista kaikkialla.
2. Tiheysfunktiolla on vain yksi huippu, joten sen maksimi on pisteessä $x = \theta$.
3. Tiheysfunktio on symmetrinen pisteen $x = \theta$ suhteen.

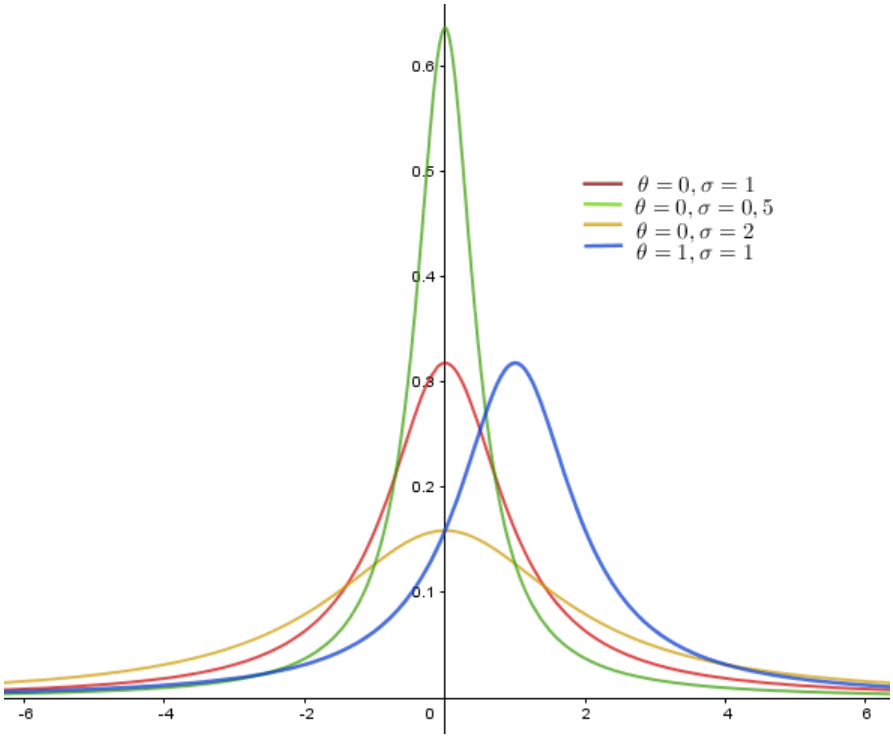
3.2 Kertymäfunktio

Yleisen Cauchyn jaukauman kertymäfunktio on

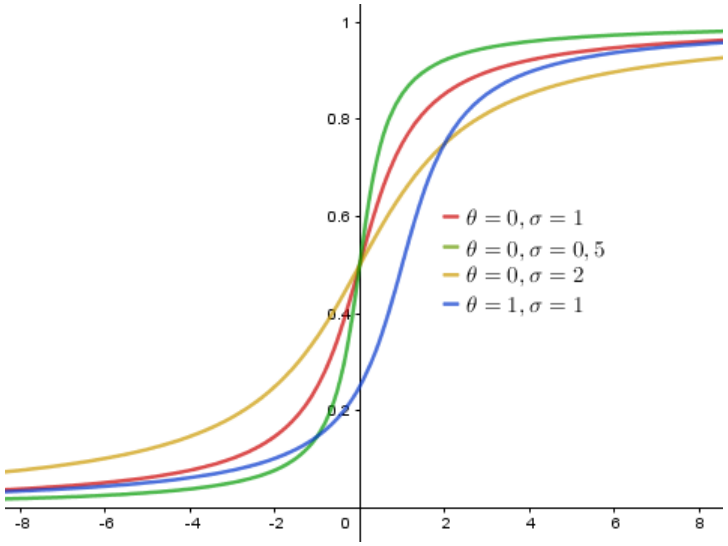
$$F(x;\theta,\sigma) = \frac{1}{\pi}\arctan\left(\frac{x-\theta}{\sigma}\right) + \frac{1}{2}. \tag{8}$$

Yleisellä Cauchyn jakaumalla ei standardin jakauman tapaan ole odotusarvoa, eikä varianssia. Kuten alussa todettiin, sillä on kuitenkin hyvin määritelty moodi ja mediaani. Standardin jakauman yhteydessä huomattiin, että k.o. tapauksessa moodi ja mediaani olivat samassa kohtaa. Yleistäen sama pätee, eli yleisen Cauchyn jakauman moodi ja mediaani ovat samat.

Seuraavalla sivulla olevat kuvat havainnollistavat tiheys- ja kertymäfunktioita eri sijainti- ja skaalaparametreilla.



Kuva 4: Cauchy-jakauman tiheysfunktioita



Kuva 5: Cauchy-jakauman kertymäfunktioita

4 Cauchyn jakauman yhteydet muihin jakaumiin

1. Jos $X, Y \sim N(0,1)$ ja X, Y ovat *riippumattomia*, niin

$$\frac{X}{Y} \sim \text{Cauchy}\,(0,1).$$

   Ts. jos X ja Y ovat riippumattomat ja normaalisti jakautuneet satunnaismuuttujat siten, että $E(X) = E(Y) = 0$ ja $\sigma_X^2 = \sigma_Y^2 = 1$, niin niiden suhde noudattaa standardia Cauchyn jakaumaa.

2. Cauchy$(0,1)$ on sama kuin t$(df = 1)$. Eli standardi Cauchyn jakauma on sama kuin Studentin t-jakauma vapausasteella 1.

3. Cauchy(θ, σ) on sama kuin $t_{(df=1)}(\theta, \sigma)$. Ts. ei-standardi Cauchyn jakauma on sama kuin ei-standardi t-jakauma.

   Studentin t-jakaumaa hyödynnetään esimerkiksi normaalijakautuneiden populaatioiden keskiarvon tutkimisessa, erityisesti kun otoskoko on pieni. Tähän emme kuitenkaan syvenny sen tarkemmin.

5 Log-Cauchyn jakauma

On malleja, joissa negatiivisten lukuarvojen ilmeneminen ei ole toivottavaa saati mahdollista. Negatiivisten arvojen esiintyminen Cauchyn jakaumassa saadaan eliminoitua käyttämällä muunneltua Cauchyn jakaumaa: log-Cauchyn jakauma. Tämän jatkuvan jakauman satunnaismuuttujan logaritmi on jakautunut Cauchyn jakauman mukaan. Tämä on normaalin Cauchyn jakauman mukaan hyvin paksuhäntäinen, sallien siten poikkeaviakin arvoja aineistoissa. Osa tutkijoista pitää sitä "super-paksuhäntäisenä", koska se on paksumpi kuin Pareto jakauma -tyypin häntä [12]. Tämän takia sitä käytetään jonkin verran tilastotieteellisissä tutkimuksissa ehdokkaana mallintamaan erilaisten aineiston tuloksia. Myöhemmin tutkielmassa esittelen pari eri tutkimusta, joissa tämä jakauma on ehdokkaana mallinnuksiin. Tällä jakaumalta puuttuu kaikki äärelliset momentit, kuten odotusarvo. Tämän todistus jätetään kuitenkin tekemättä. Loc Cauchyn jakauma on vakaa, sillä Cauchyn jakauma on vakaa. [12]

Log-Cauchyn satunnaismuuttuja on esimerkki *satunnaismuuttujan muunnoksesta*. Tällä tarkoitetaan sitä, että yhdestä satunnaismuuttujasta saadaan toinen käyttämällä apuna jotain funktiota. [24]

Määritelmä 5.1. Olkoon satunnaismuuttuja X on Cauchyn jakautunut. Tällöin sm.

$$Y = \exp(X) = e^X\,, \tag{9}$$

on log-Cauchyn jakautunut.

Lause 5.1 perustelee sen, että log-Cauhcyn satunnaismuuttuja todella on satunnaismuuttuja. Todistus lauseelle jätetään käymättä.

Lause 5.1. *Satunnaismuuttujan muunnos.* Jos X on satunnaismuuttuja ja $g : \mathbb{R} \to \mathbb{R}$ on funktio, niin myös $Y = g(X)$ on satunnaismuuttuja.

5.1 Log-Cauchyn tiheys- ja kertymäfunktion johtaminen

Johdan seuraavaksi log-Cauchyn tiheys- ja kertymäfunktiot ja lopuksi kokoan ne omiksi lauseiksi. Lauseiden todistukseksi jätetään ainoastaan funktioiden johtaminen.

Olkoon $X \sim \text{Cauchy}(\theta, \sigma)$ jatkuva satunnaismuuttuja, jonka tiheysfunktio on

$$f_X(x; \theta, \sigma) = \frac{1}{\pi\sigma}\left[\frac{\sigma^2}{(x-\theta)^2 + \sigma^2}\right], \quad \theta \in \mathbb{R}\,, \sigma > 0.$$

Tarkastellaan sm. $Y = \exp(X) = e^X$ jakaumaa. Määritetään Y:lle kertymäfunktio. Y saa arvoja $[0, \infty]$, joten selvästi $F_Y(y) = 0$, kun $y < 0$.

Kun $y > 0$, niin

$$F_Y(y) = P(Y \leq y) = P(e^X \leq y) = P(X \leq \ln(y)) = F_X(\ln(y)), \quad \text{joten}$$

$$F_Y(y; \theta, \sigma) = \frac{1}{\pi} \arctan\left(\frac{\ln(y) - \theta}{\sigma}\right) + \frac{1}{2}$$

Arvataan, että

$$\begin{aligned} f_Y(y) = F_Y'(y) &= \frac{f_X(\ln(y))}{y} \\ &= \frac{\frac{1}{\pi\sigma}\left[\frac{\sigma^2}{(\ln(y)-\theta)^2+\sigma^2}\right]}{y} \\ &= \frac{1}{y\pi}\left[\frac{\sigma}{(\ln(y)-\theta)^2+\sigma^2}\right], \qquad \text{kun } y > 0. \end{aligned}$$

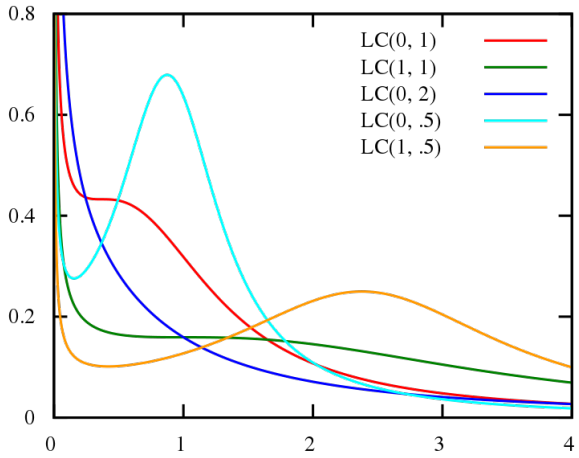
Kun $y \leq 0$, niin $f_Y(y) = 0$.

Tiheysfunktio

Lause 5.2. Olkoon satunnaismuuttuja X log-Cauchyn jakautunut. Tälläin jakauman tiheysfunktio f on

$$f(x; \theta, \sigma) = \frac{1}{x\pi}\left[\frac{\sigma}{(\ln x - \theta)^2 + \sigma^2}\right], \ \mathrm{x} > 0$$

missä $\theta \in \mathbb{R}$ on sijaintiparametri ja $\sigma > 0$ on skaalausparametri.



Kuva 6: log-Cauchyn jakauman tiheysfunktioita eri sijainti- ja skaalausparametreilla (Wikipedia).

Lause 5.3. Jos $\theta = 0$ ja $\sigma = 1$, niin tällöin log-Cauchyn jakauma on standardi log-Cauchyn jakautunut ja tiheysfunktio on muotoa

$$f(x; 0, 1) = \frac{1}{x\pi(1 + (\ln\, x)^2)}\,, \quad \mathrm{x} > 0.$$

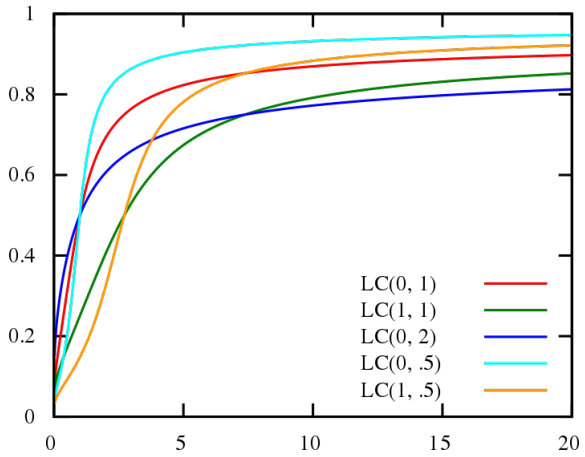
Kertymäfunktio

Kuten aiemmin huomattiin, log-Cauchyn jakautuneelle satunnaismuuttujalle X on voimassa $P(X \leq 0) = 0$, ts. tällainen satunnaismuuttuja ei voi saada negatiivisia arvoja.

Lause 5.4. Log-Cauchyn jakauman kertymäfunktio F on

$$F(x; \theta, \sigma) = \frac{1}{\pi} \arctan\left(\frac{\ln\, x - \theta}{\sigma}\right) + \frac{1}{2}\,, \quad \mathrm{x} > 0.$$

missä $\theta \in \mathbb{R}$ on sijaintiparametri ja $\sigma > 0$ on skaalausparametri.



Kuva 7: log-Cauchyn jakauman kertymäfunktioita eri sijainti- ja skaalausparametreilla (Wikipedia).

Lause 5.5. Standardin log-Cauchyn jakauman kertymäfunktio on muotoa

$$F(x; 0, 1) = \frac{1}{2} + \frac{1}{\pi} \arctan(\ln x), \quad \mathrm{x} > 0.$$

6 Cauchyn jakauman sovelluksia

Tosielämässä poikkeamia mittaustuloksissa ilmenee melko usein. Robustisten menetelmien avulla pyritään eliminoimaan näiden poikkeavien havaintojen kieleiset vaikutukset analyysituloksiin [28]. Menetelmissä käytetään paksuhäntäisiä jakaumia mallinnuksessa, sillä sellaiset sallivat aineistossa poikkeamia, toisin kuin esimerkiksi normaalijakuama. Cauchyn jakauma on paksuhäntäisyytensä takia yksi todennäköisyysjakauma, jota hyödynnetään tällaisten aineistojen mallintamisessa. Tätä jakaumaa käytetään myös mm. fysiikassa erilaisissa tutkimuksissa sekä meteorologiassa (hydrologia) sademäärien kuvaamiseen ja niiden ennustamiseen.

Tässä luvussa käsittelen kahta eri tutkimusta, joissa etsitään sopivimpia todennäköisyysjakaumia kuvamaan sademääriä *Sri Lankassa, Kolumbiassa*. Ensimmäisessä tutkimuksessa testataan kuutta eri jaukamaa, joista yhtenä on Cauchyn jakauma. Tämän tutkimuksen tavoitteena on löytää jokaiselle kuukaudelle sekä monsuunikaudelle sopiva todennäköisyysjakama mallintamaan päivän sademäärän enimmäisemääriä. Toisessa tutkimuksessa testataan 45 eri jaukaumaa, joista tavoitteena on löytää kolme sopivinta mallintamaan sademäärän enimmäismääriä sekä vuositasolla että monsuunien aikaan. Yksi näisä jakaumista on Cauchyn jakauma. Erona tutkimuksissa on siis todennäköisyysjakaumien määrät ja erilaiset ajanjaksot datan käsittelyssä.

6.1 Sademäärien jakautuminen

Erityisesti ilmanstonmuutoksen seurauksena ilmastomme on altis muutoksille. Kiinnostusta herättää äärimmäiset sääilmiöt ja niiden ilmenemistä joudutaankin todistamaan koko ajan. Sademäärien ennustamista on harjoitettu pitkään, ja ne kerrotaan ihmisille esimerkiksi säätiedotuksen yhteudessä. Tarkasti arvioituna sademäärien enimmäismäärät saattavat auttaa ehkäisemään myrskyjen ja tulvien aiheuttamia vaurioita. Erityisesti niiden alueiden, joiden sademäärät ovat hyvin suuria, sateen ennustaminen on tärkeää k.o. alueiden ihmisten kannalta: näin voidaan arvioida mahdollista tulvavaaraa ja siten taata ihmisten turvallisuus. [6]

Sademäärät vaihtelevat paljon vuoden aikana ja eri maantieteellisillä alueilla. Tämän takia on selvää, että on monia eri todennäköisyysfunktioita, joita käytetään sademäärien mallintamisessa, maantieteellisestä alueesta, sademääristä ja tutkimuksen ajanjaksoista riippen. Sademääriin liittyviä todennäköisyysjakaumia ja niiden aplikaatioita on käsitellyt useat tutkijat ympäri maailmaa [6]. Esimerkiksi Weibullin jakauma on todettu parhaimmaksi mallintamaan Japanin vuosittaista sademäärän enimmäismääriä [22].

6.1.1 Tutkimusalue ja data

Sri Lanka sijaitsee Intian valtamerellä ja on suuruudeltaan n. $65,635\,km^2$. Tämän saarivaltion ilmastoon vaikuttavat paljon topografiset piirteet, kuten tasangot, vuorijonot ja laaksot [6]. Alueella vallitsee monsuuni-ilmasto, joten vaihtuvat monsuunituulet ja niiden tuomat sateet vaihtelevat paljon vuoden aikana eri osissa saarta. Sri Lankan alueella vallitsee kaksi päämonsuunia, koillis- ja lounaismonsuunit sekä niiden väliin jäävät välikaudet. Koillismonsuunin eli *Maha-monsuunin* (joulukuu-helmikuu) aikaan tuulee koko valtiossa ja sateita esiintyy pääasiassa saaren itä- ja pohjoisosissa. Lounaismonsuuni eli *Yala-monsuuni* (huhtikuu-syyskuu) puolestaan vaikuttaa pääasiassa etelän ja lännen rannikoilla ja yälänköalueilla. Välikausien aikaan maalis-huhtikuussa, sataa pääasiassa lounaassa ja ylängöllä, kun taas loka-marraskuussa, epävakainen sää vallitsee koko maassa. [8]

Molemmissa käsittelemissäni tutkimuksissa on oletettavasti käytetty täysin samaa dataa, sillä se on molemmissa saatu meteorologian laitokselta Sri Lankasta, Kolmbiasta. Data koostuu 110 vuoden ajalta (1900-2009), jossa on koottuna valtion päivittäiset sademäärät millimetreinä (mm). Aineisto on hyvin kattava, eikä siinä ollut puutteitä minkään päivän osalta [6].

6.1.2 Ensimmäinen tutkimus

Metodologia ja tulokset

Ensimmäisen tutkimuksen *Identyfying the best probability distribution for daily maximum fainfall in Colombo, Sri Lanka* ovat kirjoittaneet Suthakaran ja kumppanit. Tässä käytetään hyödyksi seuraavia kolmea metodia parhaimpien todennäköisyysjakaumien löytämiseksi.

A. Kuvaileva tilastoanalyysi
Ennen analyysiä datasta laskettiin aineistoa havainnoittavia lukuja: keskiarvo (mean), varianssi (variance), keskihajonta (standard deviation, SD), variaatiokerroin (coefficient of variation, CV), vinous (skewness, SK), huipukkuuskerroin (coefficient of kurtosis, CK) ja sademäärän enimmäismäärä millimetreinä (mm) jokaiselle ajanjaksolle, eli joka kuukaudelle. Kuvassa 8 on koottuna tilastojen yhteenveto.

Study Period	Mean (mm)	SD (mm)	CV	SK	CK	Max
Jan	63.87	28.09	0.44	0.39	-0.56	124.7
Feb	64.24	28.44	0.44	0.65	-0.54	132.5
Mar	68.09	27.83	0.41	1.47	1.59	154.7
Apr	104.8	52.14	0.50	1.90	2.78	284.6
May	143.0	56.45	0.39	1.05	0.03	289.6
June	86.56	83.34	0.96	4.30	16.05	493.7
July	71.26	42.44	0.60	1.81	1.65	191.2
Aug	61.92	26.38	0.43	0.75	-0.62	125.9
Sep	88.31	36.08	0.41	0.29	-1.18	153.4
Oct	143.6	47.59	0.33	0.83	-0.41	256.2
Nov	122.3	48.56	0.40	1.20	0.97	270.1
Dec	76.58	22.79	0.30	0.39	-1.11	117.9
FIM	86.18	45.22	0.52	2.01	5.09	284.6
SWM	90.25	59.19	0.66	2.85	14.22	493.7
SIM	133.1	48.87	0.37	0.91	0.44	270.1
SEM	68.32	26.89	0.39	0.33	-0.50	132.5

Kuva 8: Tilastojen yhteenveto. (Suthakaran et al. [7])

B. Suurimman uskottavuuden estimointi (maximum likelihood estimation, MLE) Tämä on suosittu ja paljon käytetty tilastotieteellinen menetelmä, jonka avulla pyritään estimoimaan tilastolliselle mallille parametrit. Näiden parametrien avulla voidaan maksimoida otantadatan todennäköisyys. Tämän käyttöä varten tutkimuksessa määritetään uskottavuusfunktio, jota en sen tarkemmin tässä käsittele.

C. Tilastollisen mallin sopivuus (goodness-of-fit, GOF).
Tutkimuksessa käytettiin kolmea eri testiä, joiden avulla määritetään testijakauman ja aineiston virheitä. Pienimmän virheen tuottava jakauma voidaan todeta parhaimmaksi malliksi kyseiselle aineistolle, tässä tapauksessa tietylle kuukaudelle tai monsuunille. Seuraavien kolmen testin avulla virheitä tutkittiin: *Kolomogrov-Smirnov*, *Anderson-Darling* ja χ^2*(Chi-Square)*. Parhaiten sopiva jakauma joka

kuukaudelle valittiin pienimmän tuotetun virheen perusteella. Tutkimuksessa tuloksista todetaan, että eri kuukausien päivittäistä sademäärien enimmäismääriä kuvattaessa jakaumat vaihtelevat suuresti heittelevien sademäärien vuoksi. Cauchyn jakauma todettiin parhaimmaksi kuvaajaksi melko kuivan kuukauden, tammikuun sademäärien enimmäismäärien kuvaamiseen. GOF-testien tulosten perusteella parhaimmat jakaumat joka kuukaudelle on koottuna kuvaan 9.

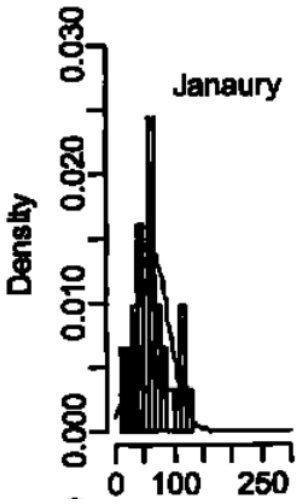
Study Period	Distribution	K-S	Chi-Square	A-D
Jan	Cauchy	0.099	0.941	0.521
Feb	LLD3	0.090	0.159	0.236
Mar	Wakeby	0.077	0.431	0.177
Apr	LP3	0.079	0.167	0.172
May	GEV	0.105	0.350	0.255
June	Wakeby	0.063	0.342	0.214
July	LLD3	0.094	1.155	0.356
Aug	GEV	0.094	2.120	0.385
Sep	GEV	0.070	0.120	0.277
Oct	LLD3	0.097	1.572	0.348
Nov	Wakeby	0.079	0.030	0.182
Dec	LP3	0.114	1.080	0.337
FIM	LLD3	0.060	2.277	0.201
SWM	LP3	0.052	0.604	0.321
SIM	LP3	0.061	2.306	0.273
NEM	Wakeby	0.058	0.234	1.579

Kuva 9: GOF-testien tulokset ja kuukautta parhaiten mallintavat jakaumat. (Suthakaran et al. [7])

Taulukkoon 1 on kirjattu Cauchyn jakauman lokaatio- ja skaalausparametrit. Kuvassa 10 havainnollisestaan sademäärien ja Cauchyn jakauman yhteensopivuutta näiden parametrien perusteella.

Kuukausi	Jakauma	Parametrit
Tammikuu	Cauchy	$\alpha = 15.54 \mid \mu = 62.21$

Taulukko 1: Parametrejä parhaiten sopivasta jakaumasta jokaisella ajanjaksolla. alfa ja myy: skaalaus- ja lokaatioparametrit



Kuva 10: Tammikuun sademäärää kuvaavan histogrammin ja Cauchyn jakauman sovitus. (Suthakaran et al. [7])

6.1.3 Toinen tutkimus

Toisen tutkimuksen *The Statistical distribution of annual maximum rainfalls in Colombo district* ovat kirjoittaneet Mayooran ja Laheetharan. Tässä tutkimuksessa aineisto jaotellaan viiteen eri aineistoluokkaan: vuosittaiseen, koillismonsuuniiin, lounaismonsuuniin ja molempiin välimonsuuneihin. Taustatiedoista ja metodologiasta kerrotaan kattavammin kuin ensimäisessa tutkimuksessa, mutta käyn läpi asiat samalla tasolla kuin aiemmassa.

Metodologia ja tulokset

Käydään läpi tutkimuksessa käytetty metodologia ja sen vaiheet.

A. Datan homogeenisyyden arviointi autokorrelaation avulla.
Tämä kertoo aineistossa ilmenevien arvojen riippuvuutta toisistaan. Autokorrelaatiota esiintyy silloin, jos sarjan uudet havainnot riippuvat edes jotenkin edellisistä havainnoista, ts. autokorrelaatiot ovat erisuuria kuin nolla. Jos havainnot eivät riipu toisistaan, autokorrelaatioiden tulee olla lähellä nollaa.[9] Tuloksen viittaavat

siihen, että viiden eri aineiston (vuosittainen, monsuunit) vuotuisten vaihteluiden kohdalla ei ole havaittavissa autokorrelaatiota.

B. Tilastollisen mallin sopivuus (goodness-of-fit, GOF).
Myös tässä tutkimuksessa käytettiin samoja GOF -testejä, kuin aiemmassakin tutkimuksessa: Kolmogorov-Smirnov, Anderson-Darling normaalisuustesti ja χ^2 -testi (Chi-square). 45:stä eri jakaumasta taulukoitiin joka aineistolle eri testien mukaan parhaiten sopiva jakauma sekä sen testin tulos (kuva 11). Cauchyn jakauma ei ollut yksi näistä. Jokaista ajanjaksoa parhaiten kuvaavat mallit koottiin yhteen taulukkoon (kuva 12).

Study Period	Test ranking first position					
	Kolmogorov-Smirnov		*Anderson-Darling*		*Chi-square*	
	Distribution	Statistic	Distribution	Statistic	Distribution	Statistic
Annual	Log-Pearson 3(3P)	0.03593	Pearson 5	0.11208	Log-logistic (3P)	0.23504
North-East Monsoon	Gen. Gamma	0.04256	Gen. Extreme Value	0.25687	Log- logistic (2P)	1.2523
First-Inter Monsoon	Burr(4P)	0.05042	Burr(4P)	0.34521	Burr(3P)	5.735
South-West Monsoon	Gen. Extreme Value	0.03222	Gen. Extreme Value	0.27369	Log-logistic (3P)	1.5995
Second-Inter Monsoon	Gen. Extreme Value	0.03514	Gen. Extreme Value	0.24044	Inv. Gaussian (2P)	0.8835

Kuva 11: GOF-testien tulokset ja jokaista ajanjaksoa parhaiten mallintavat jakaumat. (Mayooran et al. [6])

Study Period	**Best fit distribution**
Annual	Log-Pearson 3 (3P)
North-East Monsoon	Gen. Extreme Value
First-Inter Monsoon	Burr(4P)
South-West Monsoon	Gen. Extreme Value
Second-Inter Monsoon	Gen. Extreme Value

Kuva 12: Jokaista ajanjaksoa parhaiten mallintava kuvaaja. (Mayooran et al. [6])

7 Log-Cauchyn jakauman sovelluksia

Paksuhäntäisyytensä takia log-Cauchyn jakaumaa voidaan käyttää sellaisten tilanteiden mallintamiseen, jossa havainnoissa ilmenee suuresti poikeavia arvoja (outliers), esimerkiksi erilaisiin selviytymiseen liittyvissä malleissa [12].

Tässä luvussa kerron kahdesta tutkimuksesta, jossa hyödynnetään log-Cauchyn jakaumaa. Ensimmäisessä osiossa käsittelen HI-virukseen liittyvää itämis- ja tartunta-aikaa. Käsittelemässäni tutkimuksessa etsitään niitä parhaiten kuvaavia jakaumia log-Normaalista, log-logistisesta ja log-Cauchyn jakaumista. Lopputuloksena huomataan, että log-Cauchyn jakauma ei ole paras vaihtoehto tähän, mutta suurten poikkeavien arvojen sallimana on teoreettisesti kiinnostava malli. Toisessa osiossa liikutaan ekologian parissa, jossa pyritään löytämään sopiva jakauma kuvaamaan lajien runsausmallia, log-Cauchyn jakauma yhtenä vaihtoehdoista. Tutkimuksissa päädytään siihen, että log-Cauchyn jakauma on sopiva mallina tähän.

7.1 HIV

Tutkimuksen tausta

Kun ihmisen tartuttaa jokin mikrobi, sitä yleensä seuraa taudin *itämisaika* ennen kuin tautiin liittyvät mahdolliset oireet ilmenevät. Oireeton, mutta infektoitu ihminen voi välittää tautia aiheuttavaa mikrobia eteen päin huolimatta siitä, että oireita ei ole vielä ilmennyt. Tämä *tartuttamis-aika* yhdessä itämisajan kanssa vaihtelevat tartunnan saaneiden yksilöiden välillä. [10] Joissain yksilöiden kohdalla oireiden alkamiseen saattaa kulua hyvinkin pitkä aika, joka poikkeaa keskimääräisestä itämisajasta. Tällaisia poikkeavuuksia kutsutaan *poikkeaviksi arvoiksi* (outlier) [13]. Kun tällaisia poikkeavia arvoja ilmenee tutkittavassa tilanteessa, voi olla järkevää valita mallintamiseen poikkeavia arvoja salliva jakauma. Tällaiset poikkeavat arvot ovat herättäneet eriäviä mielipiteitä tutkijoiden keskuudessa siitä, ovatko ne todellasia poikkeavia tuloksia vai virheitä mittaustuloksissa [10].

Päälähteenä tässä osiossa käytän Moden ja Sleeman kirjaa *Stochastic processes in epidemiology: HIV/AIDS, other infectious diseases*. Kirjasta tarkemmin käsittelen lukua 2.4, jossa tutkitaan log-Normaalin, log-Logistisen ja log-Cauchyn jakauman sopivuutta mallintamaan HI-viruksen itämisaikaa ja tartunta-aikaa. Log-Cauchyn jakaumaa ehdotetaan itämisajan mallinnukseen juuri siksi, että se sallii suuremmatkin poikeammat aineistossa.

Tutkimuksen toteutus ja tulokset

Kun yritetään sovittaa jakaumaa mallintamaan jotain tapahtumaa, on järkevää vertailla sitä muihin jakaumiin. Kaikilla jakaumilla ei ole vertailuun sopivia parametreja, kuten odotusarvoa (Cauchy), mutta kaikille voidaan määrittää kvantiilit.

Tämän takia tutkimuksessa hyödynnetään tutkittavien funktioiden kvantiileja ja verrataan niitä keskenään. [10]

Jotta tutkittavista jakaumista saadaan jonkilainen käsitys, niiden kvantiileja verrataan yleisemmin käytettyihin jakaumiin: eksponentiaaliseen, Weibullin ja Gammajakaumaan. Nämä kolme ovat yleisesti käytettyja jakaumia erilaisssa mallinnuksissa. Tutkimuksessa kuitenkin huomautetaan, että Weibullin ja Gammajakauman oikeanpuoleinen häntä ei välttämättä ole tarpeeksi paksu, jotta ne sallisivat myös poikkeavat arvot aineistossa, kuten itämisajassa. Log-Cauchyn jakauma taas on paksuhäntäisyytensä takia hyvä kandidaatti tähän [10]. Taulukkoon 2 on listattu näiden kolmen käytetyn jakauman eri kvantiilien arvot vuosina.

q	**Eksponentiaali**	**Weibull**	**Gamma**
0.25	2.230	5.362	5.738
0.50	5.373	7.549	7.423
0.75	10.747	9.885	9.410
0.90	17.850	12.041	11.469
0.95	23.223	13.339	12.824
0.999	53.549	18.461	19.380

Taulukko 2: Valittuja kvantiileja (vuosina) jakaumille eksponentiaali, Weibull ja Gamma, samalla odotusarvolla.(Mode, C. J. and Sleeman, C. K. [10])

Log-Normaali, log-logistinen ja log-Cauchyn jakaumat riippuvat kaikki lokaatioparametrista μ ja skaalausparametrista α. Jotta voidaan vertailla tutkittavien jakaumien kvantiileja taulukon 2 arvoihin, parametrit valitaan niin, että ne noudattavat Weibullin jakauman arvoja seuaavasti: skaalausparametri valitaan siten, että mediaani on $\exp[\mu] = 7.549$ vuotta ja lokaatioparametri niin, että jokaisen tutkittavan jakauman 0.75:s kvantiili on sama kuin Weibullin jakaumalla. Taulukkoon 3 on koottu tutkittavien jakaumien, valittujen kvantiilien arvot vuosina näillä parametrien arvoilla.

q	log-Normaali	log-Logistinen	log-Cauchy
0.25	5.782	5.782	5.782
0.50	7.549	7.549	7.549
0.75	9.913	9.913	9.913
0.90	12.639	12.981	17.357
0.95	14.617	15.593	41.532
0.999	26.053	41.232	$1.409 \text{ x } 10^{38}$

Taulukko 3: Valittuja kvantiileja vuosina jakaumille log-Normaali, log-Logistinen ja log-Cauchy, samalla mediaanilla. (Mode, C. J. and Sleeman, C. K. [10])

Taulukosta 3 huomataan, että jos itämisaika noudattaisi log-Cauchyn jakaumaa, niin yksi henkilö tuhannesta saavuttaisi iän $1.409\,x\,10^{38}$ ennen AIDS:n oireiden alkamista. Ihmisen elinodotuksen nojalla ymmärretään, ettei tämä ole järkevä oletus. Siksi log-Cauchyn jakauma on epäuskottava mallina näillä parametrien arvoilla, vaikkakin teoreettisesti kiinnostava jakauma. [10]

7.2 Lajien runsausmallit

Useimmiten eliöyhteisöissä löytyy aina muutama harvinainen laji yksine edustajineen sekä muutamia yleisempiä lajeja ([18, 19, 20, 21]). Tällaista mallia eri variaatioineen on tutkittu jo vuodesta 1943 lähtien [15]. Erilaisia matemaattisia funktioita on ehdotettu lajirunauksien (SA) kuvailemisen avuksi [23]. Toinen kahdesta ehdotetusta tavasta on SA-jakaumat (SAD), joita seuraavat käsittelemäni artikkelit käsittelevät.

Lajirunsaudet ja niiden mallit ovat ekologiassa tärkeä näkökulma tutkia. Se tarjoaa paljon informaatiota esimerkiksi lajien runsaudesta, sukkessiosta ja lajien sukupuuton todennäköisyydestä, jos laji menettää elinympäristönsä. Subtrooppiset metsät tarjoavat elinympärisön suurelle määrälle lajeja, mm. erilaisille puille, joka luo hyvät mahdolliset tutkimuksiin ja ekologiseen analysointiin. [15] Alueita on tärkeä tutkia erityisesti siksi, että nämä metsät ovat kovaa vauhtia vähenemässä ja jäljelle jääneiden metsien kunto huononee [16].

Lajirunsauden malleissa on havaittavissa säännöllisyyttä: eliöyhteisössä tai taksonomisesti tai ekologisesti sukua olevien organismien välillä, kuten esimerkiksi kasveilla ja koilla metsässä [18].

Käsittelen seuraavana kolmea eri artikkelia, joissa käsitellään tutkimuksia, missä tutkitaan lajien runsautta ja pyritään löytämään runsautta parhaiten kuvaavia todennäköisyysjakaumia, eli *runsausmalleja*. Näistä ensimmäisen käsittelen hieman tarkemmin ja kaksi seuraavaa hyvin lyhyesti, vain tutkimukset ja niiden tulokset tiivistäen. Kaikissa tutkimuksissa log-Cauchyn jakauma on yksi sovitetta-

vista malleista ja jokaisessa todetaan, että se sopii runausmalliksi jopa parhaiten.

7.2.1 Lajimonimuotoisuus subtrooppisilla metsäalueilla

Ensimmäisen artikkelin *LogCauchy, log-sech and lognormal distributions of species abundance in forest communities* ovat kirjoittaneet Yin ja kumppanit. Tämän artikkelin käsittelemässä tutkimuksessa pyritään löytämään hyvä jakauma kuvaamaan puuvartisten ($\geq 1.5\,m$) lajien runsausmallia.

Tutkimusalue

Dataa kerättiin testialueilta A-E, jotka sijaitsevat subtrooppisilla metsäalueilla etelä-Kiinassa, Fengkain maakunnassa. Alueet sijaistevat 300 metriä merenpinnan yläpuolella ja ovat jokainen 1 600 m^2 kokoisia. Vertailua varten dataa kerättiin myös testialueelta F, joka sijaitsee trooppisessa metsässä lounais-Nigeriassa. Tutkimuksessa ehdotetaan lajien runsausmallien kuvaamiseen kahta uutta mallia log-normaalin (LN) sijaan: log-Cauchyn (LC) jakaumaa ja log-sech (LS) jakaumaa. [15]

Tässä tutkielmassani LS -funktion ymmärtäminen ei ole olennaista, joten jätän käsittelyn hyvin pintapuoliseksi. Tämä jakauma luetaan hyperbolisiin funktioihin. Ne ovat matemaattisia funktioita, jotka ovat määritelty eksponenttifunktion avulla. Nämä muistuttavat trigonometrisia funktioita. Hyberbolinen sekanttifunktio määritellään hyperbolisen kosinifunktion avulla. [25]

Metodologia

Lajien yksilöiden lukumäärät on jaettu ryhmiin käyttäen apuna Prestonin oktaacimetodia. Prestonin mukaan yksi oktaavi on ekvivaletti skaalauksen log_2 kanssa [20]. Tällöin R:nnes oktaavi, johon r sisältyy, esitetään intervalleissa $(2^{R-1}, 2^R)$, jossa R=0,1,2,... Jos $r = 2^R$, niin tällöin puolet lajilukumäärästä luokitellaan R:een oktaaviin $(2^{R-1}, 2^R)$ ja toinen puoli (R+1):een oktaaviin $(2^R, 2^{R+1})$ [15].

Kun havainnoidun aineiston SA-jakauma noudattaa kellonmuotoista, vasemmalta typistettyä log-normaali -jakaumaa log -skaalauksella, voidaan S* estimoida [19],[20]. S* on arvio teoreettisesta kokonaismäärästä lajeja yhdessä eliöyhteisössä. Sitä arvioidaan tutkimuksessa käyttämällä kahta erilaista metodia: summausmetodi, integraalimetodi. Summausmetodi ei log-Cauchyn jakauman kannalta ole olennainen, joten jätetään se käsittelemättä.

Integraalimetodi: kun R on jatkuva satunnaismuuttuja, niin S* arvioidaan

$$S^* \approx \int_{-\infty}^{\infty} S(R)dR = 2\int_{R_m}^{\infty} S(R)dR$$

Log-Cauchyn jakauma on jatkuva, joten integraalimetodia hyödyntäen estimoidaan S*:

$$S^* \approx \int_{-\infty}^{\infty} \frac{S_m}{1+\alpha^2(R-R_m)^2} dR = \frac{\pi S_m}{\alpha}$$

Todennäköisyysjakaumien sopivuutta aineistoihin testattiin kahden eri metodin avulla: t-testin ja Kolmogorov-Smirnov -testin (KS) avulla. Jos p-arvo < 0.05 t-testissä tai >0.05 KS-testissä, niin tällöin sanotaan, että k.o. mallin sopivuus on tilastollisesti merkitsevää.

Tulokset

Tutkimuksessa havaitaan, että kaikkien kolmen jakaumamallin sopivuus on tilastollisesti merkittävä.

(1) Määrtiyskertoimet (R_d^2) ovat kaikilla korkeat mutta korkeimmat, eli parhaimmat arvot ovat log-Cauchyn jakaumalla.

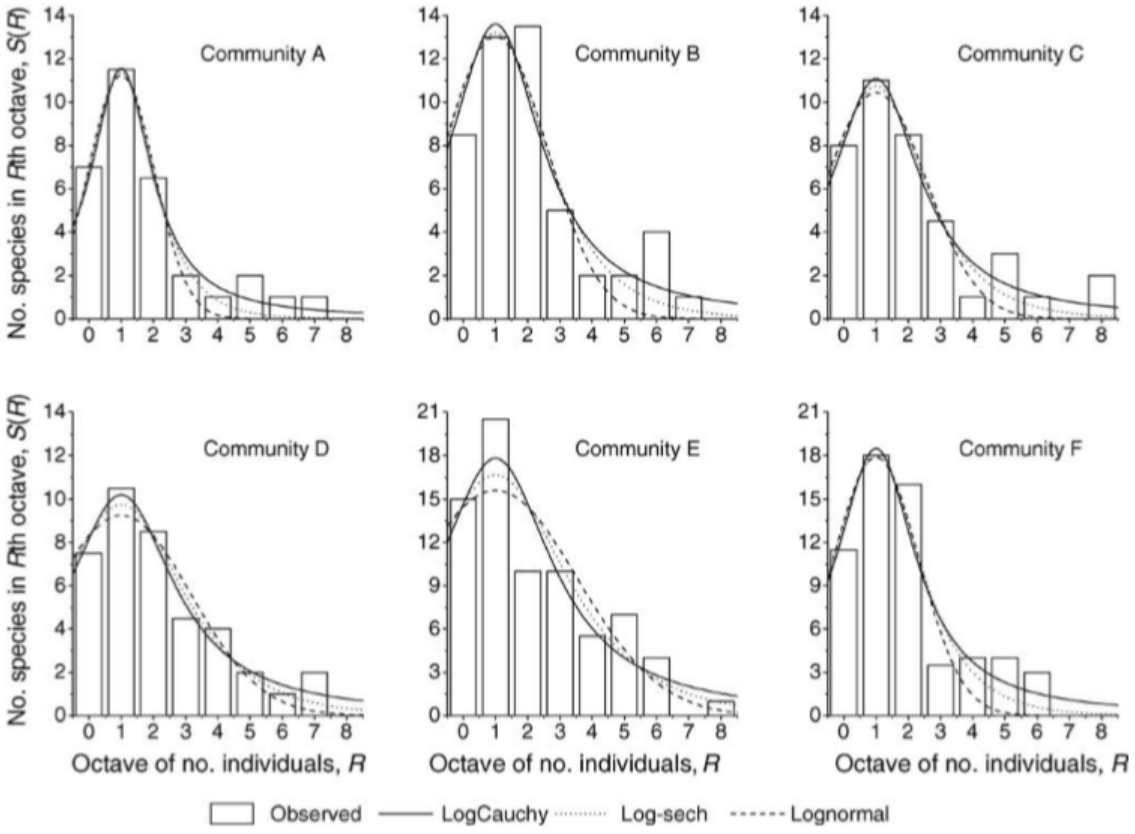
(2) t-testissä parametrien P-arvot ovat kaikilla pieniä, <0.001 jokaisen testialueen kohdalla.

(3) KS-testissä jakaumien P-arvot ovat suuria, >0.05.

Koska kaikki vasemmalta typistetyt jakaumat, erityisesti log-Cauchyn jakauma noudatti SAD:a, niin S* todella voidaan estimoida integraalimetodia apuna käyttäen. Arviot lajien määristä ovat suurempia siksi, että todennäköisesti kullakin alueella (A-F) elää enemmän lajeija, kuin niitä havaittiin (kuva 13). Kuvassa 14 tutkittavat jakaumat ovat sovitettuna aineistoihin. Kaiken kaikkiaan LS todettiin sopivan parhaiten aineistoihin ja LC seurasi toisena. Yleisesti runausmalleissa tavattava LN oli sopivuustestien mukaan kolmesta tutkittavasta jakaumasta k.o. aineistoihin huonoiten sopiva. [15]

	S	LogCauchy		
		S_{BTC}	S^*_{sum}	S^*_{int}
A	32	33.2	41.7	41.6
B	49	54.4	74.7	74.7
C	39	42.5	57.6	57.6
D	40	45.4	64.3	64.3
E	73	83.7	120.3	120.3
F	60	65.8	87.4	87.4

Kuva 13: Lajien lukumäärien vertailu alueilla A-F: havainnoidut (S), odotusarvo (S_{BTC}) vasemmalta typistetty jakaumasta summausmetodia hyödyntäen, odotusarvo (S^*_{sum}, S^*_{int}) koko jakaumasta summaus- tai integraalimetodia hyödyntäen. (Yin et al. [15])



Kuva 14: Havaitut lajirunsauden jakaumat (histogrammit) ja odotetut havainnot jakaumien log-Cauchy, log-Normaalin ja log-sech perusteella kuudella eri metsäalueella. (Yin et al. [15])

7.2.2 Lajimonimuotoisuus eri sukkession vaiheissa

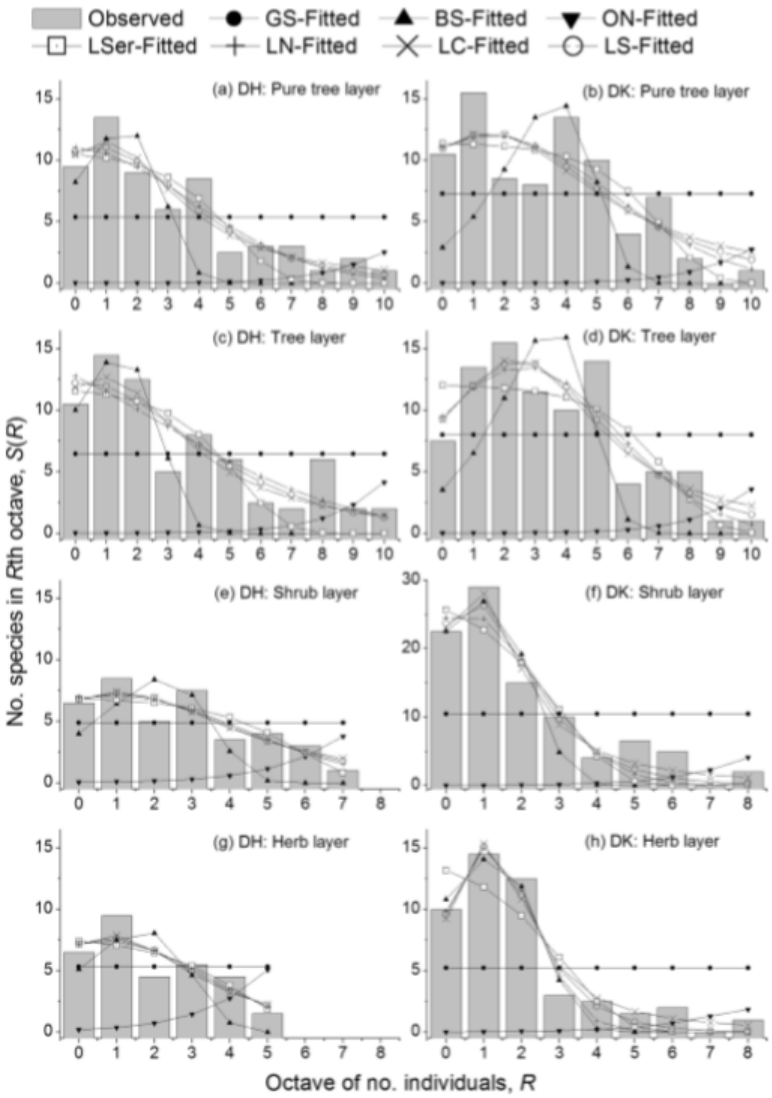
Toinen käsittelemä artikkeli on Yin:n ja kumppaneiden kirjoittama *Examining the patterns and dynamics of species abundance distributions in succession of forest communities by model selection.* Käsitellyssä tutkimuksena on tarkoituksena selvittää mm. seuraavatko eri sukkession vaiheissa olevien alueiden lajirunaudet samoja jakaumia (SAD). Keskityn tässä osiossa vain tähän tavoitteeseen, sillä muut tavoitteet ovat epäolennaisia tässä kohdassa.

Tutkimusalue

Tämän artikkelin käsittelemässä tutkimuksessa aineisto kerättiin kahdelta hehtaarin kokoiselta alueelta, jotka sijaitsevat 2 km:n etäisyydellä toisistaan, eteläKiinassa. Alueet olivat eri vaiheissa ekologista sukkessiota: toinen 70∼80 vuotta vanha ja toinen >400 vuotta vanha metsä. Sieltä kerättävät lajit olivat erilaisia putkilokasveja ja ne otettiin täysin sattumanvaraisesti jakamalla kaksi hehtaarin aluetta vielä neljään pienempään osaan, joista sattumanvaraisesti valittiin tarvittava määrä näytteitä.

Metodit ja tulokset

Molemmista kerätyistä havainnoista muodostuneet lajien runsausmallien jakaumat ovat vasemmalta typistetty siten, että niiden huippu on vasemmalla. Tämä viittaa siihen, että jokaiselta tasolta, molemilta alueilta löytyy paljon harvinaisia lajeja ja muutama yleinen laji suurine yksilömäärineen. Seitsemää käsiteltävää jakaumaa sovitettiin kaikkiin kahdeksaan eri aineistoon. Tutkimuksessa sopivuuden testaamisessa apuna käytettiin χ^2*(Chi-Square)* -testiä ja tutkittiin määrityskertoimia R_d^2. Seitsemän jakauman tuloksia vertailtiin Akaike informaatio- ja Bayesin informaatio -kriteerejä (Akaike Information Criterion, AIC; Bayesian Information Criterion, BIC) apuna käyttäen. Metodien perusteella log-Cauchyn jakauma noudatti jokaista kahdeksaa aineistoa ja sopi parhaiten seitsemään niistä.



Kuva 15: Mallien sovitus aineistoon. (Yin et al. [17])

7.2.3 Puulajien lajimonimuotoisuus

Wein ja kumppaneiden kirjoittama artikkeli *Which Models Are Appropriate For Six Subtropical Forests: Species-Area and Species-Abundance Models* käsittelee tutkimusta, jossa on kolme tavoitetta. Toisena niistä pyritään löytämään sopivat lajien runsausmallit heidän kuudelta subtrooppisen metsän alueilta saatuihin aineistoihin. Jätän kaksi muuta tavoitetta käsittelemättä, sillä ne ovat epäolennaisia tässä kohtaa.

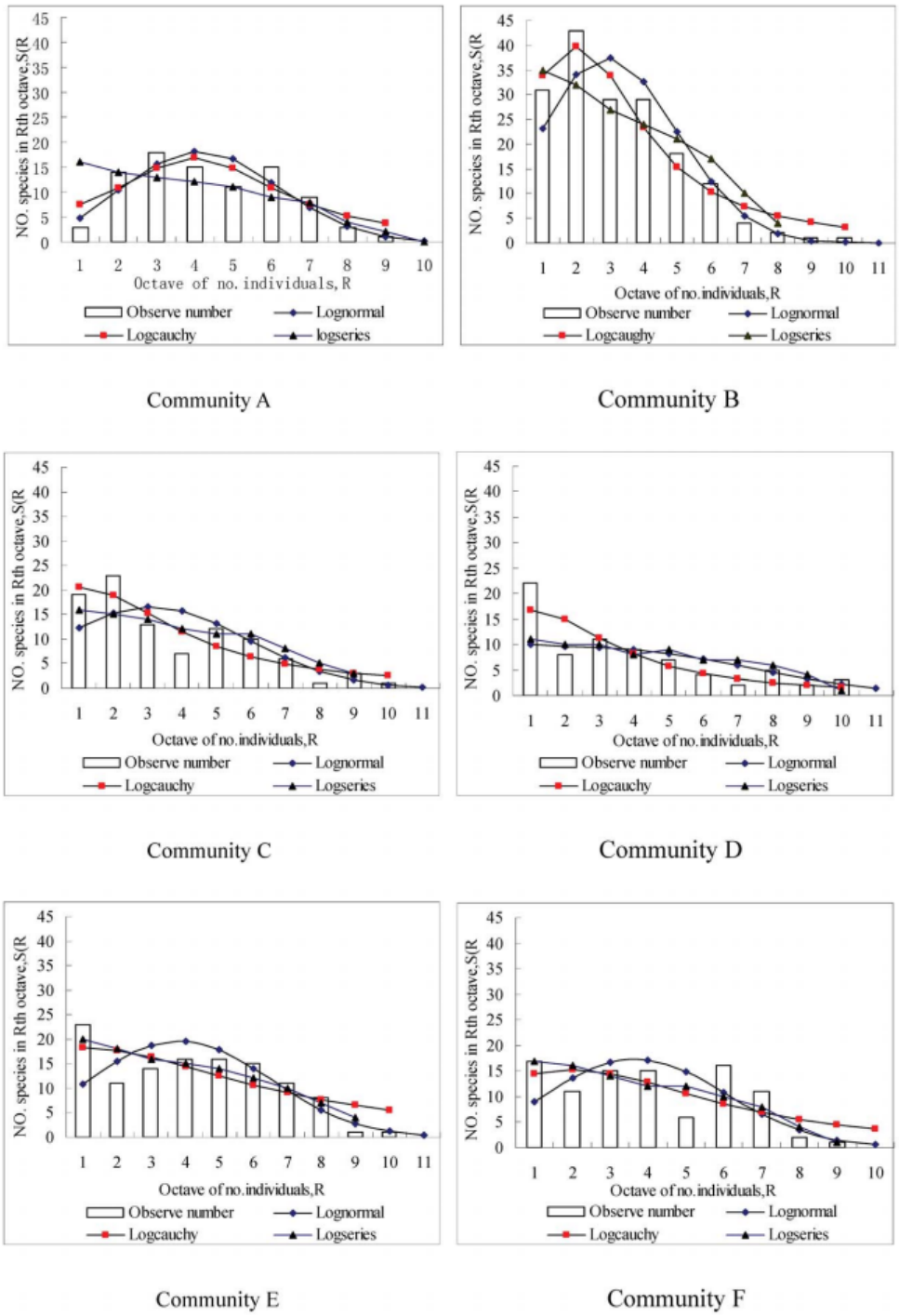
Tutkimusalue

Tutkimuksessa kerättiin aineistoa kuudesta yhden hehtaarin kokoiselta alueelta, jokainen subtrooppiselta alueelta eteläisestä Kiinasta. Lajiksi luettiin jokainen yhden senttimetrin halkaisijaltaan oleva, rinnan korkeudelle yltävä (DBH) puu. Uhanalaisija lajeja ei käytetty tutkimuksessa. Kuten aiemmissa tutkimuksissa, tässäkin aineisto jaettiin oktaavimetodia apuna hyödyntäen.

Metodit ja tulokset

Jokaisen alueen lajien kokonaismäärä estimoitiin *Estimate S*-ohjelman avulla. Käyrän sovitusta ja sopivuustestejä suoritettiin R2.3.1 -version avulla.

Puulajien monimuotoisuus hehtaarilla vaihteli suuresti näiden kuuden alueen kohdalla: havainnoidut lajit vaihtelivat 63:sta 165:een ja lajien kokonaismäärien estimaatiot 93.6:sta 229.7:ään. Määrtiyskertoimet (R_a^2) LC jakauman kohdalla olivat korkeammat kuin kahdella muulla, siten että sen (R_a^2) -arvoista kaksi on suurempaa kuin 0.8 ja kaksi suurempaa kuin 0.7. Vastaavasti seuraavaksi parhaimmat arvot sai LN, siten että kaksi oli suurempaa kuin 0.8 ja kolmantena LS, siten että yksi arvo oli suurempaa kuin 0.8 ja yksi suurempaa kuin 0.7. Kaikki kolme jakaumaa on vasemmalta typistettyjä, joka viittaa siihen että monia harvinaisia lajeja elää tutkimusalueilla, mutta niitä ei löytynyt kerätystä aineistosta. SAD:ta sovittaessa huomataan, että log-Cauchyn jakauma sopii parhaiten malliksi näihin aineistoihin kolmesta tutkittavasta jakaumasta (kuva 7).



Kuva 16: SAD mallien sovitus aineistoon. (Wei et al. [14])

8 Cauchyn jakauma lukio-opetuksessa

Cauchyn jakauma on potentiaalinen vaihtoehto erilaisten luonnonilmiöiden mallinnuksessa. Mallintaminen pitkän matematiikan opetuksessa onkin yksi esille nostettu oppimistavoite. Vaikka normaalijakauma on perinteinen ja paljon käytetty jakauma todennäköisyyslaskennassa, myös useamman jakauman käsittely kevyesti olisi mielestäni hyvödyllistä pitkän matematiikan opiskelijoille. Useamman jatkuvan jakauman käsittely voi syventää opiskelijan ymmärrystä erilaisten ilmiöiden mallintamisesta ja todennäköisyysjakaumien käytöstä niissä. Vähintäänkin käsittely syventää ymmärrystä matemaattisten kuvaajien ja jakaumien tulkinnassa sekä esimerkiksi todennäköisyydessä käsiteltävien tunnuslukujen sisäistämisessä.

Tässä luvussa pohdin Cauchyn jakauman mahdollisuuksia lukio-opetuksessa. Käyn läpi vuonna 2021 voimaan tulevaa uutta lukion opetussuunnitelmaa. Pohdin matematiikan yleisten tavoitteiden, laaja-alaisen osaamisen, oppiainerajat ylittävän opetuksen ja valinnaisen kurssin **MAA12 Analyysi ja jatkuva jakauma (2 op)** kannalta Cauhcyn jakauman sopivuutta lukion pitkän matematiikan kurssille. Koska tätä jakaumaa käytetään luonnonilmiöiden mallintamisessa, se luo hyvät puitteet lukiossa järjestettäväksi projektiluontoiseksi ja oppiainerajat ylittäväksi kurssityöksi. Ehdotukseni on siis luoda projektityö MAA12 -kurssille, joka voisi toimia yhtenä arvosteltavana työnä. Tämä lisää opettajan arvostelun monipuolisuutta, kuten arviointityössä edellytetään. Työ antaa myös opiskelijalle yhden uuden tavan osoittaa osaamistaan matematiikassa perinteisen tuntinäyttöjen ja kokeiden lisäksi. Käyn luvussa viimeiseksi läpi konkreettisen ehdotukseni projektista.

8.1 Matematiikan opetuksen tavoitteet

Opetushallitus linjaa usimmasssa luokion opetussuunnitelmassa, että matematiikan opetuksen tulee ohjata opiskelijaa ymmärtämään matemattisesti esitettyä tietoa, sen merkitystä nyky-yhteiskunnassamme ja huomata aineen välttämättömyys useilla eri aloilla, kuten tekniikassa, lääke-, talous-, yhteiskunta- ja luonnontieteellisillä aloilla sekä taiteessa. Opetuksen tulee kehittää laskemisen lisäksi luovaa ajattelua sekä erilaisten *ilmiöiden mallintamista, ennustamista* ja samalla myös ongelmanratkaisemisen taitoja.

Uusissa opetussuunnitelmissa, niin yläkoulussa kuin lukiossa, korostetaan teknologian merkitystä. Sitä tulee hyödyntää opetuksessa, mutta ennen kaikkea sen käyttöä tulee opettaa koululaisille omassa opetuksessa. Erilaisten tietokoneohjelmien käyttö nostetaan esille:

"Matematiikan opiskelussa opiskelija kehittyy hyödyntämään tietokoneohjelmistoja ja digitaalisia tiedonlähteitä oppimisessa, tutkimisessa sekä ongelmanratkai-

sussa."

Matematiikan opetuksen yleisissä tavoitteissa taas listataan, että opiskelijan tulee

"ymmärtää matematiikka sekä ainutlaatuisena itsenäisenä tieteenalana, että käyttökelpoisena välineenä, kun mallinnetaan, hallitaan tai ennustetaan yhteiskunnan, talouden tai luonnon ilmiöitä."

8.2 Laaja-alainen osaaminen ja oppiainerajat ylittävä opetus

Laaja-alainen osaaminen koostuu eri osa-aluiesta, jotka muodotavat lukion oppiaineiden yhteiset tavoitteet. Lisäksi jokaisen oppiaineen kohdalle on avattu, kuinka laaja-alainen oppiminen kyseisessä aineessa toteutuu. Laaja-alaiseen osaamiseen puolestaan kuuluu oppiainerajoja ylittävä oppiminen. Tällaisten rajoja ylittävien kokonaisuuksien hallinta ja ymmärtäminen on edellytys laaja-alaiselle osaamiselle. [26]

Matematiikan oppinaineessa laaja-alaisessa osaamisessa tuodaan ensimmäisenä esiin tärkeys tutkia arkielämän ja matematikan välisiä yhteyksiä. Oppilasta tulee kannusaa myös sinnikkääseen työskentelyyn ja ohjata pohtimaan, kuinka matematiikassa opittuja taitoja voidaan hyödyntää niin kestävässä kehityksessä kuin myös ihmiskuntaa koskevissa ongelmanratkaisuissa. Opiskelijan tulee myös saada työskennellä vaihtelevasti.

MAA12 kurssilla luonnonilmiön mallinnuksen luominen tai tutkimunen antaa hyvät edellytykset laaja-alaisten tavoitteiden saavuttamiselle. Normaalijakauma on yhden muuttujan ja kahden parametrin jakauma, kuten myös Cauchyn jakauma on. Viimeinen jaukauma on kuitenkin mielenkiintoinen, sillä sen korkeammat momentit puuttuvat ja se on hyvin paksuhäntäinen. Koska oppilaiden ei tarvitse perehtyä jakauman tutkimiseen yksityiskohtaisesti, heidän ei myöskään tarvitse opetalla uusia matemaattisia taitoja käsitelläkseen Cauchyn jakaumaa. Näin ollen matematiikan kursseilla muutenkin opiskeltavilla taidoilla opiskelija kykenee käsittelemään tätä jatkuvaa jakaumaa.

Koska matematiikassa pääpaino on itsenäisessä opiskelussa, vaihtelevuutta opiskeluun tulee järjestää kurssien aikana. Ryhmänä toteutettava projektityö luo vaihtelevuutta työskentelyyn, jota myös opetussuunnitelma edellyttää. Näin vahvistetaan tärkeitä vuorovaikutustaitoja myös matematiikassa, jotka kuuluvat ollenaisesti laaja-alaiseen osaamiseen.

8.3 Projektityö pitkän matematiikan kurssilla MAA12

Käyn läpi tässä osiossa oman ehdotuksen projektin toteuttamisesta. On tärkeä huomata, että se on suuntaa antava eikä opetuksessa testattu. Kerron luvussa lisäksi sen, miksi työ tulisi mielestäni toteuttaa juuri MAA12 kurssilla.

Valinnainen kurssi luo hyvät puitteet oppiainerajat ylittävään opetukseen, sillä jatkuvia jakaumia käytetään paljon erilaisten ilmiöiden mallinnuksessa. Tiettyä ilmiötä tarkasteltaessa mukaan voidaan helposti liittää ilmiöön liittyvät muutkin oppiaineet, esimerkiksi biologia ja maantieto. Pitkän matematiikan kurssilla **MAA8 Tilastot ja todennäköisyys** käydään läpi diskreetteihin jakaumiin liittyviä oleellisia aiheita, kuten jakaumien tärkeät tunnusluvut. Tämä helpottaa jatkuvaan jakaumaan perehtymistä.

MAA12 Analyysi ja jatkuva jakauma (2 op)

Tavoitteet
Moduulin tavoitteena on, että opiskelija

- syventää ymmärrystään analyysin peruskäsitteistä
- osaa muodostaa ja tutkia aidosti monotonisten funktioiden käänteisfunktioita
- täydentää integraalilaskennan taitojaan
- perehtyy jatkuvan todennäköisyysjakauman käsitteeseen ja oppii soveltamaan normaalijakaumaa
- osaa käyttää ohjelmistoja funktion ominaisuuksien tutkimisessa ja epäoleellisten integraalien laskemisessa sovellusten yhteydessä.

Keskeiset sisällöt

- paloittain määritelty funktio
- funktion jatkuvuuden ja derivoituvuuden tutkiminen
- jatkuvien ja derivoituvien funktioiden yleisiä ominaisuuksia
- käänteisfunktio
- funktioiden raja-arvot äärettömyydessä
- epäoleelliset integraalit
- jatkuvat jakaumat, normaalijakauma ja normittaminen

Kurssin loppupuolella keskeiset sisällöt on käsitelty, joten paljon olellaisia teemoja jatkuvista jakaumista on käyty läpi. Niiden pohjalta isompi projektityö on mahdollista toteuttaa.

8.3.1 Projektin aihe ja toteutus

Projektin aihe on sademäärien mallintaminen jatkuvien jakaumien avulla. Ideana olisi tutkia normaalijakauman lisäksi Cauchyn jakauman sopivuutta Suomen sademäärien mallintamisessa. Vertailun ja pohdinnan tueksi oppilaiden tulee kerätä tutkimuksia sademäärien mallinnuksesta myös muualta maailmasta.

Projektin koonti ja toteuttaminen vaatii kenties aikaa muutaman oppitunnin verran. Ajatus olisi, että ryhmä mittaa kuukauden ajan jokaisen päivän sademäärät. Koulu tietenkin suorittaa mittaukset haluamallaan tavalla ja valitsee sopivan aikavälin tähän. Voidaan pohtia, hoitavat oppilaat itse mittaukset vai käyttääkö ryhmä opettajan keräämiä tuloksia. Jos mittaukset vievät paljon aikaa, seuraavina vuosina voidaan käyttää jo valmiiksi kerättyjä tuloksia. Jokainen oppilasryhmä kokoaa taulukoksi kuukauden sademäärät. Näiden tulosten avulla lasketaan tarvittavat tunnusluvut, kuten sademäärän keskiarvo ja mediaani. Tuloksista tulee luoda histogrammi, kun taas Cauchyn jakauman voi piirtää esimerkiksi Geogebran avulla. Jakauman ja histogrammin sovitukseen voidaan käyttää valittua matemaattista ohjelmaa.

Myös eri maantieteellisiltä alueilta olisi hyvä saada tutkimustuloksia. Tällöin omien tulosten ymmärtäminen ja vertailu helpottuu ja lopun pohdinta saadaan kattavaksi. Matematiikan pohdintaosuudessa opiskelijoiden tulee miettiä, kuvaako Cauchyn jakauma Suomen päivittäisiä sademääriä paremmin kuin normaalijakauma. Loppupohdintaan voidaan lisätä muutakin podittavaa aikataulusta riippen, esimerkiksi: onko Cauchyn jakauma todettu hyväksi malliksi sademäärien kuvaamisessa jossain muualla?

Kuinka rajat ylittävä oppiminen toteutuu projektissa?

Projektiin on helppoa ja järkevää liittää muita oppiaineita mukaan, aineenopettajien yhteistyön sallimissa rajoissa. Maantiedon oppitunneilla voidaan esimerkiksi pohtia topograafisten piirteiden vaikutusta sademääriin niin Suomessa kuin muualla maailmassa. Oppilaat voivat selvittää, minkälaiset sademäärät ovat tyypillisiä esimerkiksi monsuunien aikaan tietyillä maantieteellisillä alueilla. Biologian tunneilla voidaan miettiä, mitä hyötyä ja haittaa sateesta on ja millasistä määristä kasvit hyötyvät. Esimerkiksi osa kasveista on sopeutunut hyvin kosteisiin oloihin ja vaativat runsaita sateita. Tiedonhaku on olennainen osa tulosten vertailussa, joten mukaan projektiin saadaan liitettyä myös äidinkieli. Nämä kaikki oheistutkimukset voidaan suorittaa k.o. aineen oppitunneilla.

Tällaisen projektin kokonaisuuden hahmottaminen ja ymmärtäminen kehittää laaja-alaista ja oppiainerajat ylittävää osaamista. Lisäksi arvioitavana työnä oppilaan osaamisen arviointia saadaan monipuolisestettua. Kaiken kaikkiaan tällä kokonaisuudella saadaan toteutettua opetussuunnitelman määrittelemiä tavoitteita niin oppilaan oppimisessa kuin myös opettajan arviointityössä.

Viitteet

[1] Tuominen Pekka, *Todenäköisyyslaskenta 1*, Limes ry, 10. muuttumaton painos, 2010.

[2] Linde Werner, *Probability Theory : A First Course in Probability Theory and Statistics*, De Gruyter, 2017.

[3] Wikipedia, Cauchy distribution, `https://en.wikipedia.org/wiki/Cauchy_distribution`. Viitattu 14.4.2020.

[4] Wikipedia, Arkustangentti ja kotangentti, `http://www.math.jyu.fi/matpo/kirja/rfa/index-148.html#pgfId-49286`. Viitattu 14.4.2020.

[5] Stephanie, Statistics How To, June 22,2015, `https://www.statisticshowto.datasciencecentral.com/cauchy-distribution-2/`. Viitattu 24.10.2018.

[6] Mayooran, T., Laheetharan, A., *The Statistical distribution of annual maximum rainfalls in Colombo district*, Department of mathematics and statistics, University of Jaffna, Jaffna, Sri Lanka. Sri Lankan journal of applied statistics, Vol (15-2). Accepted 2014.

[7] Suthakaran, R., Perera, K., Wikramanayake, N., *Identyfying the best probability distribution for daily maximum fainfall in Colombo, Sri Lanka*, Conference proceedings - International forum for mathematical modeling 2014.

[8] `https://www.aasiamatkat.fi/blogi/sri-lanka-paras-matkustusajankohta.htm`. Viitattu 14.4.2020.

[9] Box, G. E. P., and Jenkins, G. (1976), Time Series Analysis: Forecasting and Control, Holden-Day.

[10] Mode, C. J. and Sleeman, C. K. (2000). Stochastic processes in epidemiology: HIV/AIDS, other infectious diseases. World Scientific.

[11] Wikipedia, Incubation period, `https://en.wikipedia.org/wiki/Incubation_period`. Viitattu 14.4.2020.

[12] Wikipedia, Log-Cauchy distribution, `https://en.wikipedia.org/wiki/Log-Cauchy_distribution`. Viitattu 14.4.2020

[13] Wikipedia, Outliers, `https://en.wikipedia.org/wiki/Outlier`. Viitattu 14.4.2020.

[14] Wei SG, Li L, Chen ZC, Lian JY, Lin GJ, Huang ZL, Yin ZY. (2014). *Which Models Are Appropriate For Six Subtropical Forests: Species-Area and Species-Abundance Models*

[15] Yin ZY, Peng SL, Ren H, Guo QF, Chen ZH. (2004). *LogCauchy, log-sech and lognormal distributions of species abundance in forest communities.* Ecol Model 184: 329-340.

[16] Laurance WF. (2007). *Forest destruction in tropical Asia.* Curr Sci India 93: 1544-1550.

[17] Yin ZY, Zeng L, Luo SM, Chen P, He X, Guo W, Li B. (2018). *Examining the patterns and dynamics of species abundance distributions in succession of forest communities by model selection.*

[18] Molles, M.C., 1999. *Ecology: Concepts and Applications.* McGraw-Hill, New York, 509 pp.

[19] Fisher, R.A., Corbet, A.S., Williams, C.B., 1943. *The relation between the number of species and the number of individuals in a random sample from animal population. J. Anim. Ecol. 12, 42-58.*

[20] Preston, F.W., 1948. *The commonness and rarity of species.* Ecology 29, 254-283.

[21] Preston, F.W., 1962. *The canonical distribution of commonness and rarity.* Ecology 61, 88-97.

[22] Hirose, H. 1994. *Parameter Estimation in the Extreme-Value distributions using the Continuation Method.* Information Processing Society of Japan, Vol 35, 9.

[23] Harte, J., Kinzig, A., Green, J., 1999. *Self-similarity in the distribution and abundance of species.* Science 284, 334-336.

[24] Hyvönen, V., Lauha, P., Tolonen, T., 2018. *Todennäköisyyslaskenta 1.* Helsingin yliopisto, luentomoniste.

[25] Wikipedia, Hyperbolinen funktio, `https://fi.wikipedia.org/wiki/Hyperbolinen_funktio`. Viitattu 14.4.2020.

[26] Aalto, S., 2019. *Pareto-jakauma ja sen sovelluksia sekä mahdollisuuksia opetuksessa* . Helsingin yliopisto, pro gradu -tutkielma.

[27] Opetushallitus. (2019). Lukion opetussuunnitelman perusteet 2019. https://www.oph.fi/sites/default/files/documents/lukion_opetussuunnitelman_perusteet_2019.pdf Viitattu 16.4.2020.

[28] Suomen Pankki, 2004. *Tutkimus aikasarjojen poikkeavien havaintojen vaikutusten vähentämisestä ARIMA-aikasarjamallien täsmentämisessä.* Tiedote nro 4. https://www.suomenpankki.fi/fi/media-ja-julkaisut/tiedotteet/2004/tutkimus-aikasarjojen-poikkeavien-havaintojen-vaikutusten-vahentamisesta-arima Viitattu 16.4.2020.